DESIGN MODELS

Design models accommodate qualitative, categorical variables. We typically refer to these as models of analysis of variance. However, in the GLM framework, these qualitative variables (categories with no numeric value) are considered factors. But, there is some utility in considering these models as design models for our purpose, to understand the role of a design matrix and consider the possible operations of design matrix in matrix algebra.

Consider a model for three means, $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_3$. The mean for group 1 (perhaps school 1) can be considered to be a function of the grand mean (μ, the population mean), plus the unique effect of group 1 (α_1) and the residual for group 1 (e_1).

$$\bar{y}_1 = \mu + \alpha_1 + e_1$$
$$\bar{y}_2 = \mu + \alpha_2 + e_2$$
$$\bar{y}_3 = \mu + \alpha_3 + e_3 \quad \rightarrow \quad \bar{y}_i = \mu + \alpha_i + e_i$$

The general form of these models for the three means can be combined into:

$$\bar{y}_i = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e_i$$

And the three models can be represented in design matrix notation as a set of indicators for the group membership as:

$$\begin{array}{llll} \bar{y}_1 = 1 & 1 & 0 & 0 \\ \bar{y}_2 = 1 & 0 & 1 & 0 \\ \bar{y}_3 = 1 & 0 & 0 & 1 \end{array}$$

Notice the first column represents the grand mean (constant or intercept), and each subsequent column represents a corresponding group membership, where an individual is in group 1, group 2, or group 3. Each row (for a group mean) can be read as: The group 1 mean is a function of the grand mean (coded as 1) and membership in group 1 (coded as 1) and not being a member of group 2 (coded as 0) or group 3 (coded as 0).

The full matrix notation to represent this system of equations should be familiar to you:

$$\begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

$\underline{\bar{y}} = \mathbf{X}\,\underline{\alpha} + \underline{e}$ where $\mathbf{X}$ is the design matrix

We need to solve for $\underline{\alpha}$ group effects from the equation

$$\underline{\bar{y}} = \mathbf{X}\,\underline{\alpha} + \underline{e}$$

Remember we need to be able to take the inverse of $\mathbf{X}$ but it is not square. So we use the method we found in regression and we can solve for the group effects, the $\underline{\alpha}$ values, as:

$$\underline{\hat{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\underline{y}.$$

This is exactly the same function we use to solve for the betas in regression.

Take a close look at the design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

There is linear dependence in the columns. Notice, column 1 = column 2 + column 3 + column 4. Our $\mathbf{X}$ is singular.

In this case, $\underline{\hat{\alpha}} \neq (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\underline{y}$ since $|\mathbf{X}'\mathbf{X}| = 0$

In all design models, $|\mathbf{X}'\mathbf{X}| = 0$; design models are of deficient rank.

To solve this problem, we need to "reparameterize" the design matrix.

Example:

		α: Type of Twins	
		Monozygotic	Dizygotic
β: Gender	Male	$\bar{y}_{11}$	$\bar{y}_{12}$
	Female	$\bar{y}_{21}$	$\bar{y}_{22}$

Here we have a model with two factors, gender and type of twins. There is some unknown dependent variable (maybe achievement or personality measure).

This is an additive main class model; containing only main effects. Notice the design matrix contains the constant or grand mean and then the four combinations of gender and type of twin.

$$\begin{bmatrix} \bar{y}_{11} \\ \bar{y}_{12} \\ \bar{y}_{21} \\ \bar{y}_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \underline{e}$$

$$\underline{\bar{y}} = \mathbf{A}\,\underline{\xi} + \underline{e}$$

This is the typical representation of such models, where $\underline{\xi}$ (ksi) is the parameter vector. Since $\mathbf{A}$ is of deficient rank, we cannot invert the matrix to solve the equations.

We can reparameterize the model to solve the deficiency problem.

Add equations:
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} = D \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} \rightarrow \begin{matrix} 0 = \alpha_1 + \alpha_2 \\ 0 = \beta_1 + \beta_2 \end{matrix}$$

Which suggests that $\Sigma\alpha_i = 0$ & $\Sigma\beta_i = 0$; ANOVA assumptions. The sum of the gender effects and the sum of the twin-type effects both are zero. That means once we estimate the effect of being male, we don't need to estimate the effect of being female because it is redundant, since they sum to zero.

So by augmenting the design matrix with these two equations, we do not change the results of the original model – we only add a constraint on the group effects; that they sum to zero. This is NECESSARY to make the matrix positive definite.

This results in:

$$\begin{bmatrix} \underline{y}. \\ \cdots \\ \underline{0} \end{bmatrix} = \begin{bmatrix} A \\ \cdots \\ D \end{bmatrix} \underline{\xi} + \begin{bmatrix} \underline{e} \\ \cdots \\ \underline{0} \end{bmatrix}$$

Notice there is a new symbol, $\cdots$, which denotes augmentation of a vector or a matrix, like appending two matrices together. Now we have a solution for $\underline{\xi}$ (the parameter vector, which hasn't changed). We can now solve for the parameters of the model. Notice we have the same result. Let's call the augmented matrix AD; so we have: AD-transpose AD inverse, AD-transpose y.

$$\underline{\hat{\xi}} = \left(\begin{bmatrix} A \\ \cdots \\ D \end{bmatrix}' \begin{bmatrix} A \\ \cdots \\ D \end{bmatrix} \right)^{-1} \begin{bmatrix} A \\ \cdots \\ D \end{bmatrix}' \begin{bmatrix} \underline{y}. \\ \cdots \\ \underline{0} \end{bmatrix}$$

This may not be the "best" solution because the introduction of $\underline{0}$ is arbitrary, although this is the classical ANOVA method.

The ideal method of reparameterization involves careful selection of the augmenting matrix. This allows us to test other parameters simultaneously as a component of the design. There is great benefit to considering how to parameterize the design matrix, because we can build in *a priori* contrasts or other interesting statistical hypotheses. The alternative is to test for interesting group differences through *post-hoc* comparisons, which has the tendency to increase Type-I error rates, since there are multiple tests being conducted as though they are independent. If we build them into the design matrix, we can do these more formally through *a priori* or planned statistical testing, which is more of a confirmatory approach, rather than exploratory and *ad hoc*.

Consider our original design model, where the design matrix is $\mathbf{A}$ and the parameter vector is $\underline{\xi}$. This can be generalized to the case of *n* (number of unique groups) × *m* (number of parameters):

$\underline{\bar{y}} = \mathbf{A}\,\underline{\xi} + \underline{e}$ → with dimensions: $(n\times 1) = (n\times m)(m\times 1)+(n\times 1)$, where n = # of cells in design

Let $\mathbf{A}$ $(n\times m)$ be of rank *l*, where $l \leq m$.

This is deficient rank because in the design matrix, there will always be linear dependence as a function of the grand-mean.

Now we can define an augmentation matrix $\mathbf{L}$.

Let **L** be of rank *l* and linearly independent on rows of **A**.

$$\text{Rank}\begin{bmatrix} A \\ \cdots \\ L \end{bmatrix} = \text{Rank}(\mathbf{A}) = \text{Rank}(\mathbf{L}) = l$$

The key to the whole process is choosing **L**. We can choose **L** based on research hypotheses (*a priori* contrasts among groups). The best reparameterization for solving equations involving a design matrix includes a set of *a priori* contrasts.

Consider a one-way design (one factor) with 4-levels (four groups):

$$\underline{\bar{y}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \underline{e}$$

Examples of many possible **L**s includes the following:

Simple Contrasts: $\quad L = \begin{bmatrix} 1 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} \mu + \dfrac{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}{4} \\ \alpha_1 - \alpha_4 \\ \alpha_2 - \alpha_4 \\ \alpha_3 - \alpha_4 \end{bmatrix}$

The first row provides the grand mean (since the sum of the group effects is zero); the other three simple contrasts compare each group to group 4.

Helmert Contrasts $\quad L = \begin{bmatrix} 1 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1 & -1/3 & -1/3 & -1/3 \\ 0 & 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} \cdots \\ \alpha_1 - \dfrac{\alpha_2 + \alpha_3 + \alpha_4}{3} \\ \alpha_2 - \dfrac{\alpha_3 + \alpha_4}{2} \\ \alpha_3 - \alpha_4 \end{bmatrix}$

Helmert contrasts are also orthogonal contrasts.

You can also create contrasts to assess your own specific research comparisons or a priori comparisons.